Word Frequencies in English-Language Literature, 1700-1922

Genre-specific word counts for 178,381 volumes from the HathiTrust Digital Library [v.0.1]

Note that this is a **beta** data release. Please send feedback to tunder@illinois.edu.

Many of the questions scholars want to ask about large collections of text can be posed using simplified representations – for instance, a list of the words in each volume, together with their frequencies. This dataset represents a first attempt to provide that information for English-language fiction, drama, and poetry published between 1700 and 1922, and contained in the HathiTrust Digital Library.

The project combines two sources of information. The word counts themselves come from the HathiTrust Research Center (HTRC), which has tabulated them at the page level in 4.8 million public-domain volumes. Information about genre comes from a parallel project led by Ted Underwood, and supported by the National Endowment for the Humanities and the American Council of Learned Societies. This project applied machine learning to recognize genre at the page level in 854,476 English-language volumes. Mapping genre at the page level is important because genres are almost always mixed within volumes. Volumes of poetry can have long nonfiction introductions; volumes of fiction can be followed by many pages of publishers' advertisements. Fortunately, text categories of this broad kind (fiction /nonfiction/poetry/drama/paratext) can be identified fairly accurately by statistical models.

Because both of these projects mapped volumes at the page level, their results can be collated to produce genre-specific datasets. Researchers who want access to the richest level of detail can do this for themselves, pairing the data and metadata described above in order to extract individual pages that belong to a given genre. To provide an easier point of access, we have done that pairing for you here, and aggregated pages at the volume level. So the fiction dataset below, for instance, contains only volumes that were identified as containing fiction, and in each case aggregates the wordcounts for *only the pages that were identified as fiction*.

To provide an even easier point of access, we have also created yearly summary files for each genre. But before relying on those summaries, consult "provenance of the volumes" below. The yearly summaries include everything published in a given year, which may not be the cross-section you want.

Contents

volumes of fiction	101,948
volumes of poetry	58,724
volumes of drama	17,709

Resources

- Spelling normalization and OCR correction
- Methods used for genre prediction
- More information about HathiTrust datasets

Attribution

Ted Underwood, Boris Capitanu, Peter Organisciak, Sayan Bhattacharyya, Loretta Auvil, Colleen Fallaw, J. Stephen Downie (2015). *Word Frequencies in English-Language Literature, 1700-1922 (0.2) [Dataset].* HathiTrust Research Center. doi:10.13012/J8JW8BSJ.

Provenance of the volumes; methods of selection

To use these datasets, literary historians will need to understand what kind of sample they represent. First of all, this is a selection of nonserial volumes published between 1700 and 1922, and it should include only volumes whose primary language is English. (Volumes in other languages were filtered out using both library metadata and automatic language recognition. But the dataset certainly does include literature in translation.) Volumes bear dates of publication, not dates of composition, so some volumes (particularly of poetry and drama) will be reprints of works written several centuries earlier.

Secondly, the dataset includes only volumes held by large public or university libraries that contribute to HathiTrust. This is a diverse and large collection, but some categories of publication, such as gift books or dime novels, may not be preserved in proportion to their original circulation. Contributing institutions are mainly located in the United States. So while the collection contains volumes from around the globe, coverage of works published in the US is more complete. Also, because books before 1800 may be held closely in Special Collections, digitization of that period is less predictable. We don't necessarily recommend this dataset as a source for literary research before 1750.

Finally, the volumes and pages aggregated here have passed through the filter of automatic genre recognition, and represent a particular model of genre. In practice, this model should align quite closely with readers' intuitions. For a full description of the methods used to produce page-level genre predictions, see "Understanding Genre in a Collection of a Million Volumes." A short version: the model's predictions about genre matched human descriptions 93.6% of the time, which is roughly as often as our six human readers agreed with each other (94.5%). Moreover, the datasets provided here have passed through additional (automatic and manual) filtering that allows us to guarantee better than 97% precision. (In other words, fewer than 3% of the pages in these collections come from a genre other than the one listed). However, there is no such thing as a perfect model. Genre boundaries are inherently blurry. For instance, while we've tended to categorize "closet dramas in verse" as drama and "dramatic monologues" as poetry, in reality there are shades of gray between those categories. The dataset presented here also deliberately sacrifices some recall for precision: it's likely that roughly 17% of the fiction and 30% of the poetry pages in HathiTrust are left out of this sample. (Much of this loss took place when we filtered out generically heterogenous collections to improve precision: scholars who want a recall rate closer to 90% can consult our original, unfiltered page metadata.)

Even with the limitations described above, we believe that this dataset can help distant readers provide a fuller picture of literary culture than currently available. But we don't necessarily mean that all scholars will want to use this entire dataset. We anticipate that most scholars who use this resource will want to construct a corpus by sampling or selecting some subset of these volumes, rebalanced to reflect the particular cross-section of literary culture they are attempting to understand. For some questions, for instance, it will be appropriate to include reprints, or literature in translation, or books for a juvenile audience; for other purposes, those volumes should be excluded. We haven't made those decisions; the construction of an appropriate sample is ultimately up to individual researchers, and claims about historical representativeness can only be made relative to their particular research goals.

Data format; other technical details

To make download easier, volumes have been grouped in tar.gz files characterized by genre and publication date. Each of these files will unzip into a folder containing separate files for each volume. Each filename (minus the .tsv extension) is a HathiTrust volume ID; these IDs can be paired to the summary metadata table we provide in order to identify author, title, date, and so on. (More importantly, they can be used to retrieve full metadata from the HathiTrust, and to view the original text online.) Each volume-level file contains two tab-separated columns: the first column contains words (or other tokens) and the second column indicates how often that token appears in the volume. We count only occurrences in pages that were identified as belonging to the relevant genre; running headers and footers are also excluded from the count, because a book that has "Three Musketeers" on the top of every page would otherwise end up with an implausible number of musketeers.

To normalize frequencies, researchers will also need a count of all words or all tokens counted. This can be created simply by summing the counts for the volume. For some purposes, you may want to limit this total to alphabetic tokens, or to words recognized by an English dictionary. (The latter option can be a simple way to compensate for the reduction in counts created by imperfect OCR.)

The data provided here is drawn directly from the HTRC's Extracted Features dataset v.0.2, and reflects decisions about word boundaries and tokenization made as part of that extraction process. In order to provide data of manageable size and complexity, we have simplified the original structure by collapsing different parts of speech. For instance, we add together occurrences of "rose" (noun) and "rose" (verb) to produce a single count for the grapheme. We have also normalized all tokens to lower-case, whereas the original dataset is case-sensitive. Scholars who need the distinctions we removed can, of course, retrieve the original data.

Resources for normalization and OCR correction

Optical character recognition makes errors. A constant level of randomly-distributed error is not a significant obstacle to research, but in practice OCR errors are not distributed randomly; they tend to cluster in cheaply-printed books, and especially in books printed before 1820. Scholars pursuing research across that temporal boundary may want to make an effort to minimize error. Resources to facilitate that process are provided in Resources above; for instance, CorrectionRules.txt is a translation table that can be used to correct common OCR errors.

In most cases, correction is straightforward at the token level; it's easy to see, for instance, that "whieh" is an error for "which." However, there are a special group of cases for which correction and normalization become difficult after context has been discarded. It's difficult to know whether "fame" is an error for long-s "same" unless you can see context like "that fame day." We wanted to support corrections of this kind without imposing our corrections on all users. So we have provided correction files for each genre that indicate recommended additions and subtractions for these special ambiguous words. These recommendations are based on Ted Underwood's process of contextual OCR correction (which used the original texts). Corrected results will still be imperfect; the goal is only to mitigate and reduce error. In correction files, the first column contains a volume ID, the second column contains a word, the third column contains a positive or negative integer that should be added to the count for that word to reflect contextual correction. These files also propose recommended corrections for certain word-division inconsistencies that would be hard to detect without context (e.g., "any where" becomes "anywhere" around the middle of the nineteenth century). Here, technically, we are overstepping the line of mere "correction" and proposing "normalization" of the dataset, since "any where" is the dominant/correct spelling before 1840.

The correction files we provide only include changes that require context to discern; we are providing this information because it would otherwise be impossible for researchers to infer it. But there is also a much larger set of corrections that could be made to individual tokens. Decisions about normalization can be just as important, if you want to compare works from different periods or different sides of the Atlantic. For instance, "honour" and "honor," "today" and "to-day," are still distinct in this dataset. If you want to treat them as equivalent, see Resources for lists of variant spellings and unevenly-hyphenated forms, as well as a list of OCR correction rules.

Yearly summaries

To provide the gentlest possible on-ramp to this data, we have also aggregated the volumes for each genre in a yearly summary file. This provides a table that can be easily manipulated in R or even Excel, but of course forecloses the option of selecting/excluding particular volumes. We have adjusted these yearly counts using all the correction resources described above (normalizing spelling and hyphenation as well as correcting OCR using contextual and single-token methods). We have also divided uncommon hyphenated words at the hyphen and restricted the dataset to the 10,000 most common tokens in each genre. The resulting tables are easy to use, but it's important to remember that they include reprints of texts originally published much earlier.

Because we only include the 10,000 most common words in the yearly summary file, scholars can't necessarily reconstruct a total wordcount for each year by summing up all the words. Instead, we have provided three special tokens for each year: `#ALLTOKENS` counts all the tokens in each year, including numbers and punctuation; `#ALPHABETIC` only counts alphabetic tokens; and `#DICTIONARYWORD` counts all the tokens that were found in an English dictionary.

Metadata

The metadata of record for this collection is stored in the HathiTrust Digital Library; see their bibliographic API. To make access easier, we have provided summary metadata tables for each genre. This information has been extracted from MARC records using a custom script, xmlparser.py. Most of the columns will be self-explanatory; some that aren't are explained on p. 40 of "Understanding Genre." It's particularly important to know that "publication date" is recorded in a complicated way in controlfield 008 of MARC (because, for instance, books may list a range of years rather than a single one). We have provided the original MARC information (in four columns), but have also provided a single column called "date" that records our inference about the earliest likely date. This is the column we have used to organize volumes in tar files. The "subjects" column contains information extracted from multiple parts of the MARC record -- including genre as well as subject information, and some information from controlfield 008 of MARC -- but, frankly, you may want to take all of it with a grain of salt. Cataloging for books from this period is often incomplete or unreliable.

Links to specific files

For each genre, we provide a metadata file, a corrections file, and a yearly summary, as well as tar.gz files that aggregate individual volume-level wordcount files, sorted by estimated date of publication. Metadata files and yearly summaries are small (less than 30MB); some of the tar.gz files can be larger (up to 550 MB).

Fiction

fiction_metadata.csv

fiction_yearly_summary.csv

fiction_contextual_corrections.csv

fiction_1700-1799.tar.gz

fiction_1800-1834.tar.gz

Poetry

poetry_metadata.csv

poetry_yearly_summary.csv

poetry_contextual_corrections.csv

poetry_1700-1799.tar.gz

poetry_1800-1834.tar.gz

Drama

drama_metadata.csv

drama_yearly_summary.csv

drama_contextual_corrections.csv

drama_1700-1799.tar.gz

drama_1800-1834.tar.gz

fiction_1835-1869.tar.gz	poetry_1835-1869.tar.gz	drama_1835-1869.tar.gz
fiction_1870-1879.tar.gz	poetry_1870-1879.tar.gz	drama_1870-1879.tar.gz
fiction_1880-1889.tar.gz	poetry_1880-1889.tar.gz	drama_1880-1889.tar.gz
fiction_1890-1894.tar.gz	poetry_1890-1894.tar.gz	drama_1890-1894.tar.gz
fiction_1895-1899.tar.gz	poetry_1895-1899.tar.gz	drama_1895-1899.tar.gz
fiction_1900-1904.tar.gz	poetry_1900-1904.tar.gz	drama_1900-1904.tar.gz
fiction_1905-1909.tar.gz	poetry_1905-1909.tar.gz	drama_1905-1909.tar.gz
fiction_1910-1914.tar.gz	poetry_1910-1914.tar.gz	drama_1910-1914.tar.gz
fiction_1915-1919.tar.gz	poetry_1915-1919.tar.gz	drama_1915-1919.tar.gz
fiction_1920-1922.tar.gz	poetry_1920-1922.tar.gz	drama_1920-1922.tar.gz

Acknowledgements

Primary funding for the "Understanding Genre" project was provided by the National Endowment for the Humanities and by the American Council of Learned Societies. Computation was supported by the Institute for Computing in Humanities, Arts, and Social Sciences at the University of Illinois (I-CHASS).

Any views, findings, conclusions, or recommendations expressed in this release do not necessarily represent those of the funding agencies.

Shawn Ballard, Michael L. Black, Jonathan Cheng, Nicole Moore, Clara Mount, and Lea Potter also deserve acknowledgment for their work on the "Understanding Genre" project, particularly in the creation of page-level training data.